图书情报工作
第63卷 第10期 2019年5月

科研用户访问国外学术数据库的隐私保护与对策*

艾琼[1] 刘纯璐[2] 游林[3]
[1] 杭州电子科技大学图书馆 杭州 310018 [2]杭州电子科技大学通信工程学院 杭州 310018
[3] 杭州电子科技大学网络空间安全学院 杭州 310018

摘要:[**目的/意义**]针对我国科研工作者获取国外学术信息过程中涉及的隐私保护问题,提出相关建议、对策与防范策略,并探讨大数据与区块链隐私保护技术及其实现隐私保护的可行策略。[**方法/过程**]通过对国外学术数据库平台网站的访问数据进行收集分析,发现这些数据库平台除了在其自身网站中插入收集用户信息的脚本外,还存在使用第三方平台脚本对用户信息进行实时收集的情况,并对此结果进行分析,挖掘其关联性。[**结果/结论**]通过大数据技术,科研用户访问国外学术数据库信息可为数据库商及其合作伙伴或政府机构所获取利用,致使科研用户的隐私信息发生泄露的风险,从而导致个人安全问题、财务问题,甚至国家安全问题,需要从多个层面加以重视。

关键词: 大数据 国外学术数据库 科研用户 隐私保护 区块链
分类号: G203
DOI:10.13266/j.issn.0252-3116.2019.10.002

在互联网背景下,用户对信息的获取从纸本文献转向数字资源,从简单的原文获取转向更深层次的信息挖掘和数据分析。为了应对这一变化,承载数字资源的学术数据库平台通过对用户个人及其数据检索活动的信息进行收集、分析和预测,了解用户的阅读爱好与研究兴趣,掌握用户的需求动态,从而有针对性地为每个用户提供个性化的推送服务,进而更广泛、更牢固地吸引用户利用其数据库资源[1]。然而,在大数据环境下,通过大数据技术对科研用户信息进行越来越活跃的收集、比较、聚类、分析和交易,也必然会导致用户隐私信息被泄露的现象发生[2]。

当今,由于人们的许多工作与生活活动都处于网络环境下,个人信息大都被各种网络系统(如电子邮件系统、社交平台、数据库网站、电商平台等)所收集保存。2017年3月央视曝光了有超过50亿条公民信息泄露。涉案信息主要来自京东电商以及多家知名互联网公司的用户。2018年3月17日,美国《纽约时报》和英国《卫报》共同发布报道,剑桥分析公司(Cambridge Analytica)在未经用户本人同意的情况下收集了Facebook的5 000多万用户的个人信息,用于在2016年美国总统大选期间有针对性地推送广告,影响大选结果。这些用户信息是以"学术研究"之名而被盗用的。此后据证实Facebook又发生了8万多个用户账户的私人信息被盗,并被试图以每个账户10美分的价格在网站上出售。后Facebook方面称被盗数据很可能是通过恶意浏览器插件而获取的。欧洲议会主席A. Tajani称,"Facebook用户数据的滥用是不可接受的对公民隐私权的侵犯"[3]。

我国科研用户获取国外学术信息主要依赖于对国外各种相应的学术数据库平台进行检索。这些数据库资源平台来自不同数据库商家或学术机构,除了少量在国内设立镜像站提供服务外,绝大多数都是在境外提供网络信息服务。因此,国内科研用户在获取国外学术资源信息的过程中,必然存在着个人隐私信息遭泄露的风险隐患。一个国家的科研用户的整体信息情况,密切关系到该国的综合科技水平,乃至国家综合生产力情况与综合国力情况,甚至会泄露该国的国防实力。因此,学术数据库的科研用户的个人隐私信息安

* 本文系浙江省教育厅科研项目"大数据下高校数字图书馆的云存储与服务安全问题研究"(项目编号:Y201636443)和浙江省教育技术研究规划课题"基于云环境下海量数字资源云服务辅助教学研究"(项目编号:JB015)研究成果之一。
作者简介:艾琼(ORCID:0000-0001-8230-5294),馆员,E-mail:aiq@hdu.edu.cn;刘纯璐(ORCID:0000-0002-9876-7685),硕士研究生;游林(ORCID:0000-0002-1383-0500),教授。
收稿日期:2018-09-27 修回日期:2018-11-22 本文起止页码:12-20 本文责任编辑:刘远颖

全问题，也极大地影响到国家的网络信息安全。目前虽然没有报道有关科研人员的个人信息被某学术数据库泄露的新闻消息，但其实大部分科研人员的个人信息早已被某些学术数据库泄露了，只不过是由于目前没有发生对科研人员个人或相关机构或国家产生较大危害的事件，也或者由于一些安全因素的原因，对产生的危害不便于向外界公开。如笔者在没有注册的情况下，就经常收到一些学术数据库或网站的电子邮件（如 ResearchGate、Academia.edu、Researcher Academy）以及一些从来不曾关注过的国外学术期刊的电子邮件。从这些电子邮件可以看出，他们几乎都对笔者的研究领域与科研兴趣很了解，也知晓笔者的一些其他个人信息。

图书馆在为科研用户提供学术数据库资源服务时，可考虑利用结合现代密码算法的大数据隐私保护技术或云服务安全存储技术等，对涉及用户个人隐私及其涉密研究领域的信息等采取一定的安全保护措施，防止信息的泄露与滥用。

1 国外学术数据库平台网站嵌入用户信息收集脚本及分析

1.1 数据库网站嵌入脚本概况

数据库平台要对用户的个人及行为信息进行收集，必须借助网站服务才能实现，其中在网站源代码中插入的各种脚本程序是获取网络访问日志收集用户访问信息的主要手段。为了追踪这一过程，笔者利用火狐浏览器 Firebug 插件对国外数据库平台网站的网络访问数据进行了实时捕捉分析，发现各个数据库平台网站都嵌入了各种脚本程序，除了平台自身网络日志获取脚本外，还使用了其他网站分析商、数据提供商、广告服务商等来自第三方平台的信息收集脚本。笔者在 2017 年 4 月收集了相关数据，并重点选择了 5 家国外数据库平台（见表 1），分析其脚本嵌入的情况。这 5 家数据库平台分别是在电气电子工程、计算机科学、人工智能、机器人、自动化控制、遥感和核工程领域收录期刊影响因子和被引用量都名列前茅的 IEL 数据库（IEEE/IET Electronic Library），全球最大的科技与医学文献出版发行商 Elsevier 旗下的 ScienceDirect 数据库，专注科学技术、数学以及医学领域内具有较高影响力的 SpringerLink 数据库，全球工程研究领域知名的文摘数据库工程索引（EI）以及在图书馆电子书领域知名的 OverDrive 电子书平台。此外，笔者还依据清华大学图书馆常用外文数据库列表[4]，选取 30 家数据库抓取了其网络访问数据（见表 2）。

表 1 与表 2 数据显示，Google、Adobe、Facebook 这些网络服务巨头都出现在脚本来源里，它们占据着大数据服务的主要市场。此外，一些新创建的大数据创业公司如 Usabilla、Redlink、Pendo 等，与学术数据库平台的自身收集脚本一起织就了捕捉科研用户及其活动信息的一张大数据之网。30 家数据库中，除 EBSCO 数据库与 Web of Science 平台只采用自身的脚本外，其他数据库都存在嵌入第三方平台信息收集脚本的情况。其中使用 Google 相关服务的最多，达到 26 家，有 15 家网站嵌入第三方的脚本来源超过 2 个。另外，在收集数据过程中还发现，数据库网站也会动态调整其收集脚本。

表 1　5 家国外数据库平台网站嵌入第三方信息收集脚本概况

数据库名称	脚本域	脚本来源	来源平台属性
IEL 数据库	coremetrics.com	IBM	网站分析和营销优化工具
	g.doubleclick.net	Google	谷歌广告服务
	googletagservices.com	Google	谷歌标签管理服务
ScienceDirect 数据库	adobedtm.com	Adobe	DTM 分析工具
	google-analytics.com	Google	谷歌数据统计服务
	usabilla.com	Usabilla	用户分析服务
	googletagservices.com	Google	谷歌标签管理服务
SpringerLink 数据库	g.doubleclick.net	Google	谷歌广告服务
	google-analytics.com	Google	谷歌数据统计服务
	z.moatads.com	Moat	Moat 分析服务
EI 数据库	pendo.engineeringvillage.com	Pendo	Pendo 分析服务
	adobedtm.com	Adobe	DTM 分析工具
OverDrive 电子书平台	google-analytics.com	Google	谷歌数据统计服务
	g.doubleclick.net	Google	谷歌广告服务

表 2 30 家外文数据库平台网站嵌入第三方信息收集脚本一览

	数据库名称* \ 脚本来源	Google	Adobe	Usabilla	Twitter	Facebook	Redlink	Webtrends	Pendo	Coreme-trics
1	ACM	√		√						
2	ACS	√			√					
3	AGU	√	√							
4	AIAA	√					√			
5	AIP	√				√	√			
6	AMS	√								
7	APS	√								
8	ASCE	√								
9	ASME	√								
10	Elsevier (SD)		√	√						
11	Emerald	√						√		
12	EV (EI)		√						√	
13	HighWire	√								
14	IEL (IEEE/IET)	√								√
15	IMechE	√		√						
16	Jane's	√								
17	Nature	√								
18	OSA	√								
19	ProQuest	√								
20	RSC	√								
21	SAE	√	√							
22	SAGE Journals	√		√						
23	Science Online	√	√		√	√				
24	SIAM	√			√					
25	SPIE	√								
26	SpringerLink	√								
27	Taylor & Francis	√								
28	Wiley	√	√							
29	EBSCO									
30	Web of Science									
	总计	26	6	4	3	2	2	1	1	1

* 数据库名称对照：ACM(美国计算机协会)、ACS(美国化学学会)、AGU(美国地球物理学会)、AIAA(美国航空航天学会)、AIP(美国物理联合会)、AMS(美国数学学会)、APS(美国物理学会)、ASCE(美国土木工程师学会)、ASME(美国机械工程师协会)、Elsevier((爱思唯尔出版社 ScienceDirect 数据库)、EV(EI，工程索引)、HighWire(美国斯坦福大学图书馆创立的开放获取学术数据库)、IEL(IEEE/IET，美国电气电子工程师学会/英国工程技术学会)、IMechE(英国机械工程师学会)、Jane's(简氏军事装备技术及国家安全数据库)、Nature(《自然》杂志数据库)、OSA(美国光学学会)、ProQuest(ProQuest 公司创立的数据库检索平台)、RSC(英国皇家化学学会)、SAE(美国汽车工程师协会)、SAGE Journals(SAGE 出版社期刊数据库)、Science Online(《科学》杂志数据库)、SIAM(美国工业与应用数学会)、SPIE(国际光学工程学会)、SpringerLink (施普林格期刊与电子书数据库)、Taylor & Francis(Taylor & Francis 出版社期刊数据库)、EBSCO(EBSCO 检索平台)、Web of Science(Web of Science 检索平台)

1.2 嵌入脚本功能与网站隐私政策分析

数据库平台在服务进程以及运营管理过程中，采集到的用户相关信息的质量、数量和准确性，决定着该数据库平台公司决策的科学性。所以，为了增强平台的综合实力和竞争力，平台在全面对用户信息进行收集的同时，还会采取各种途径追踪用户信息。平台将已收集的用户信息与从其他平台收集到的用户信息通过大数据技术进行整理、分析和聚合后，就可获得更多潜在的有效信息来优化其管理或运营决策[5]。通过浏览各家学术数据库平台网站的隐私条款，就可以一窥网站收集的信息内容及其使用方式，更好地理解嵌入脚本的功能。

作为全球最大的专业学术组织，IEL 在学术研究领域发挥着重要作用的同时，也极大地影响着全球用户的学术研究工作。IEEE 在其隐私政策[6]中提及，IEL 会利用线上、线下的注册和表单，通信和互动、销售查询和交易记录来全面收集个人信息，其中就包括联系方式、交易信息、用户偏好、金融/信用卡和付款信息、活动和会议参与资料，以及用户使用服务和产品的统计数据。另外，IEL 会使用 Cookie、像素标签和其他技术来帮助第三方供应商在网页中投放和追踪广告。

ScienceDirect 数据库是 Elsevier 公司数字化产品平台，EI 也是其旗下收购的重要文摘类数据库。Elsevier 隐私政策[7]中明确指出，用户信息会与 Elsevi-

艾琼，刘纯璐，游林.科研用户访问国外学术数据库的隐私保护与对策[J].图书情报工作,2019,63(10):12-20.

er全球公司集团共享。Elsevier公司在自己平台收集用户信息的同时,会与所有关联公司合作来获得更全面的用户信息,从而更科学地优化公司决策来获得盈利。

Springer在其数据库平台SpringerLink的隐私政策[8]中声明:除了通过注册表单来收集用户数据为其提供友好、高效、安全的互联网服务外,还将利用Cookie和Google Analytics(谷歌数据统计与分析服务)来收集信息,用户使用过程中所产生的信息会被发送并存储在谷歌公司在美国的服务器上。

OverDrive的一般隐私政策[9]则更加详细具体化,直接声明"为了提高服务水平和提升客户体验,需要收集更全面的用户信息来针对性地分析每个客户的需求",其收集的信息基本是全方位覆盖的。所收集的信息不仅包括个人信息,还包括在线活动、数字内容选择、评论、评分、读者证号、Internet协议地址、设备类型以及唯一设备数据等信息。平台收集用户个人信息和其他相关信息的目的包括:支持服务内部范围更广的操作;向用户发送其可能感兴趣的产品信息以及新服务功能的信息;提供个性化服务来提高用户体验;通过匿名策略将用户个人信息与若干第三方(如广告商、业务伙伴等)分享。此外其还明确指出"会利用Google Analytics以及其他软件来全方位地跟踪网站用户的个人活动信息"。

总之,国外学术数据库平台都存在利用网站嵌入脚本收集用户个人账户信息及其活动信息的情况,并可能与第三方分享这些收集到的信息。另外,数据库平台及数据分析服务提供商也存在为了遵守法律、法庭命令、诉讼或其他法律程序,响应来自内部或用户的居住国以外的政府当局要求,而透露收集到的用户信息。如果平台认为披露是合理必要的,为了执行其权利或保护其经营,平台也可能会透露用户的个人隐私信息。虽然各平台网站都声称会保护用户的隐私信息,但在这种形势下,科研用户显然处于一种被动的状态,其个人隐私信息有可能被滥用或泄露。

2 科研工作者访问国外数据库时隐私信息泄露与安全问题

我国广大科研工作者在从事科研工作时,往往需要大量地查阅文献资料。在当今数字化时代,科研资料基本都以数字化形式存储在数据库平台中,以便科研工作者能非常便捷地进行查阅与下载。我国目前虽也建立了中国知网(CNKI)、维普中文科技期刊数据库、万方数据知识服务平台、中国科学引文数据库等知名学术数据库。但目前科学技术研究方面的大量优质文献仍是集中在国外知名的如Ei Village、Web of Science、ScienceDirect、IEL、Wiley、SpringerLink、ACM等各种类型的数据库中,所以科研工作者在进行科研活动的过程中经常要访问这些国外数据库,查阅或参考相关的科研资料。

科研人员在获取这些国外数据库中的学术信息的过程中,其个人信息大多时候都会在默认同意的情况下被悄无声息地收集。当用户个人信息被全天候、全方位地收集并被分享利用或发生泄露时,就有可能损害到用户的个人隐私。同时,由于每个科研用户的背后都对应着其工作的高新技术企业、科研单位、高等院校或国家重要部门,因此当个人和相应机构的信息被全面收集与分析利用时,就不仅可能产生个人隐私泄露的风险,而且也可能给科研机构乃至国家的重要保密信息带来安全问题。

对于个人用户,学术数据库平台通过收集用户信息来挖掘潜在需求,从而为用户提供个性化服务,但在这个过程中也无形地打造了"数字全景监狱",使隐私信息很大程度上在用户无意识的状态下被服务商有意识地收集、分析和传播。学术平台根据用户提交的注册表单可以获得读者的年龄、学历、性别、定位和银行卡号等个人信息,还可以通过对用户访问网站过程中的操作步骤、实时鼠标点击状态的收集来获取用户使用数据库的习惯和行为偏好。1939年美国制定的《美国图书馆协会道德规范》就指出图书馆用户的私人信息应被视为机密[10]。而这其中的私人信息就被指明包括个人、组织机构等实体不愿意被外部知道的信息,比如个人的行为模式、位置信息、兴趣爱好、健康状况、公司财务状况等[11]。

数据挖掘是指利用一定的计算机算法或方法从大量的数据中搜索或析出隐藏于其中信息的过程。通常的数据挖掘算法类型有统计方法、在线分析方法、情报检索方法、机器学习算法、专家系统和模式识别算法等或这些方法的组合利用。大数据分析一般是指利用计算机算法、系统或统计方法等,对收集来的大量数据进行检查、分析,以发现隐性的、未知相关性的、市场趋势性的、客户偏好性的或其他有用信息并加以详细研究和概括总结的过程。采用数据挖掘技术与大数据分析技术相结合策略,可以从用户借阅或下载文献的历史数据中析出用户的个人兴趣嗜好、阅读偏好、研究方向、政治倾向、身体状况等敏感信息。将用户信息属性

项进行关联分析也能得到更深层、更全面的信息[12]。将收集到的出现在互联网上的个人信息与数据库平台中公开的研究数据以及系统设定的敏感关键词进行关联并展开大数据技术分析，就有可能了解用户当前的个人研究方向、内容与进展，并可以进行下一步的预测。当这些来自数据库平台的大量用户数据与其他第三方社交平台（如 ResearchGate、Facebook 等）的数据相结合，也就可以更深入地了解与甄别单个用户的生活状况、学术圈及其正在进行的研究课题内容与现状，甚至可暴露出用户重要的个人隐私信息，如所从事的研究方向领域、研究课题内容、个人职业与职位、家庭信息，甚至身份证件等信息。国内万方数据的知识服务平台脉络分析系统，中国知网的知识元检索、CNKI 学术趋势、学者圈都是在此基础上进行的初步尝试。

当一个机构引进某个学术数据库服务后，按照数据库采购流程，一般都需要提供机构的网络地址 IP 段用于识别合法的用户服务请求。而当属于该机构中的所有个人信息、机构相关信息被收集分析后，就可以了解该机构的主要研究领域方向、内容、进展、创新和技术需求等信息。由于个人、机构与国家是密切相关的，所以当一个国家若干科研人员的科研活动及其密切相关的科研工作进展状况都被一个有敌对行为的机构或国家有意识地广泛收集、聚合和分析并被有目的利用时，这将不再只是一个简单的个人隐私问题，有可能上升到国家层面的网络信息安全问题，乃至国家安全问题。

3 科研用户获取国外学术信息的隐私保护对策

《中国网民信息安全状况研究报告》显示，绝大多数网民对网络信息安全处于无助状态，并且有近 90% 的网民在个人信息遭泄露后无任何补救措施。科研用户从利用互联网访问学术数据库网站开始，其个人数据就在不断地被平台收集获取。要保护用户的个人隐私信息，图书馆可经常性地对科研用户开展个人隐私保护方面的讲座与宣传。或对图书馆数字资源登陆系统进行改进，增设插件功能来实现提醒用户注意个人隐私信息，帮助用户树立其个人隐私信息的保护意识。在进行文献资料检索时尽量使用安全的网络数据检索、查询与下载工具。在政策层面上也应该设立或委托专门的机构来对引进购买的国外数据库平台进行信息安全评估与监管，设立数据库安全等级评估制度，同时在法规与技术策略上规范数据库信息采集的权限。

3.1 树立隐私保护意识

随着“互联网 +”技术的快速发展，电商平台、社交平台以及电子支付系统等已日益深入人们的工作与生活中。个人手机号、E-mail 地址、身份证号、住址、银行帐号、家庭收支情况等信息常常被泄露于互联网上或被不法分子盗用，并常给人们的工作与生活造成较大的损失或伤害。依据百度与 Google 检索，近几年来不断发生支付宝或银行帐号被盗而导致资金被非法转移的事件，这在很大程度上是由于人们的个人隐私保护意识不强，对自己个人信息的管理缺乏安全防范观念、账户的密码设置安全强度不高，存在“撞库”和“扫号”等攻击的风险。有许多科研用户的研究领域常涉及到国家的高新前沿科技甚至国防上的尖端机密技术，他们的个人隐私信息的泄露极有可能会潜在地给国家安全或国防建设事业造成不可估量的危害，目前这样的案例已发生过多次。

因此，在高等学校、科研院所、图书馆乃至全社会范围内，开展个人隐私的网络保护普及教育，让广大科研用户接受隐私保护理论与实务培训极为必要。通过宣传、培训等途径让科研用户了解各大学术数据库或科研交流平台的隐私保护政策，提升用户在获取学术信息过程中的隐私保护意识。在不影响学术交流的情况下，提醒科研用户注意个人信息的留存。虽然从 2017 年 6 月 1 日开始施行的《中华人民共和国网络安全法》对保护我国公民的个人信息进行了的严格规定，但对在境外进行违法盗取并利用我国公民个人信息的行为在监管与打击上都存在很大的难度。

3.2 使用安全的网络工具

在用户访问数据库平台网站时，网络浏览器总是兢兢业业地记录着用户的所有操作。虽然互联网访问方便了用户对感兴趣信息的了解与获取，但信息的无选择性的收集极有可能损害到用户的隐私安全。有多家网络/信息安全机构表示，黑客等不法分子可以通过各种网络爬虫或抓包等软件对用户的浏览行为及内容进行数据采集，恶意获取用户的个人信息。因此建议科研用户使用安全的浏览器来实现安全的网络访问。目前，Chrome、Firefox、Safari、IE 等浏览器都纷纷推出隐身浏览（Private Browsing）功能，其本质就是阻止网站跟踪用户行为，使用户在使用这些浏览器时，他们的浏览行为及内容不会被跟踪及记录下来，或不会与用户的身份信息关联起来。另外，也可以考虑为科研用户设计专门的具有隐私保护功能的浏览器，提供安全访问国外学术数据库的网络途径，从而实现科研用户能

艾琼，刘纯璐，游林. 科研用户访问国外学术数据库的隐私保护与对策[J]. 图书情报工作,2019,63(10):12－20.

以匿名化身份访问数据库网络,使他们的个人隐私信息得到安全保障。

3.3 设立专门的评估与监管机构

目前曝光的因个人信息泄露而造成的重大刑事案例,基本都不是因用户个人原因而造成的信息泄露,而都是因服务产品系统存在漏洞或服务商管理存在问题而导致的,所以需要从国家立法角度来约束服务机构。2016 年 12 月,国家互联网信息办公室发布《国家网络空间安全战略》报告[13],指出“建立实施网络安全审查制度,加强供应链安全管理,对党政机关、重点行业采购使用的重要信息技术产品和服务开展安全审查,提高产品和服务的安全性和可控性,防止产品服务提供者和其他组织利用信息技术优势实施不正当竞争或损害用户利益。”科研数据库的应用涉及到国家的科学研究现状,关系着国家的科技创新与未来发展趋势,甚至涉及到国防科技的机密研究领域。因此,图书馆在引进国外的数据库时可以考虑设立或委托专门的机构对数据库及其相应的使用平台的隐私保护策略进行评估与监管。对不符合我国个人信息安全等级保护规范的学术数据库可要求其进行相应的安全优化。

2006 年 5 月开始实施的《信息安全技术数据库管理系统安全评估准则》对一般的数据库管理系统的信息安全等级的评估要求进行了规范,但其中没有从用户的角度对隐私信息的保护做任何规范要求。所以期望今后国家图书馆机构以及全国信息安全标准化技术委员会,能依据《中华人民共和国网络安全法》,建立针对我国科研工作者的隐私保护政策或保护法规,制定相应的国内及引进国外学术数据库的评估标准,同时建立相应的保护科研用户隐私监管机制。

3.4 建立涉及用户隐私信息的学术数据库信息安全等级

在引进学术数据库的过程中,由专门的评估机构为数据库设立不同的信息安全等级。检索重要的数据库的科研用户一般都具有较高的学术水平及较高的学位或职称,所以重要的数据库的使用与管理应该更重视用户信息的保护。图书馆等提供数字资源服务的机构应在采购这类数据库的合同条款中规定数据库系统需要对登陆用户采取较严格的隐私保护措施。不言而喻,对数据库建立安全等级也就自然提醒用户,同时用户在使用数据库时应该考虑使用更安全的网络工具。

2006 年 12 月开始实施的信息安全技术国家标准《信息安全技术 数据库管理系统安全技术要求》对数据库管理系统的安全技术作了 5 个等级的技术要求:

(1)第一级是用户自主保护级。主要是使用户具备自主保护的能力,如对用户的身份进行鉴别,对用户进行访问控制,对用户的真实身份信息与注册信息进行安全保护。

(2)第二级是系统审计保护级。主要是指数据库系统应具有能进行自主安全审计的功能,能实现对访问数据库的用户身份与数据、访问行为(日期与时间、登陆数据库的次数、登陆成功与否)以及数据库是否能免遭篡改或破坏等进行审计,并提供审计日志。其宗旨是强调用户对自己的访问行为负责的保护策略。

(3)第三级是安全标记保护级。主要是指数据库应提供有关安全策略模型、数据标记、主体对客体强制访问控制的非形式化描述。凡需进入数据库管理系统的用户,应先进行标识或建立账号,应对登陆到数据库管理系统的用户身份的真实性进行鉴别。这种保护级是监督性的安全保护。

(4)第四级是结构化保护级。主要是指数据库系统应支持建立于一个明确定义的形式化安全策略模型,强调要求将第三级安全保护级中的自主和强制访问控制扩展到所有主体与客体。这其实是一种更强制性的保护策略要求。

(5)第五级是访问验证保护级。主要是指访问数据库需要满足访问监控器需求。访问监控器仲裁主体的全部访问。访问监控器本身是抗篡改的,且必须足够小,能够进行分析和测试。

数据库管理系统安全就是要求不仅数据库中存储的数据信息进行安全保护,而且要求对用户的注册信息进行保护。在此安全技术要求中,最核心的就是强调对用户的访问控制以及用户身份的鉴别。在目前的数据库管理系统中,身份鉴别一般是采用口令鉴别或基于令牌(如手机令牌)的动态口令鉴别。但现实中,基于时效性,通常都是采用口令鉴别的方式。为了安全性,口令除了在选取时应包含字母、数字以及一些符号外,一般需要作加密处理后再保存,以防止穷举搜索攻击及内部管理人员的泄露。依据目前的信息安全技术现状,对口令的加密一般用对称密码算法 AES,或者用 Hash 函数 SHA-1 对口令做摘要,或者采用 AES 加 SHA-1 对口令进行双重保护。但由于 2017 年 2 月荷兰国家数学和计算机中心(CWI)和 Google 的研究人员找到了 SHA-1 的碰撞实例,证明了 SHA-1 的不安全性。所以建议今后在进行口令鉴别时采用 SHA-2 来替代 SHA-1。

生物特征鉴别技术已成为各种数字管理系统的用

户身份鉴别方法。但是使用这种方法，在注册时需要对用户的具有唯一特征的生物特征进行采样获取模板后生成用户的生物特征模板库。而在目前情况下，这个模板库的生物特征模板都没有经过任何形式的加密处理或采取其他安全保护措施，所以模板库极易被系统管理员泄露或遭到黑客攻击。而一旦用户的生物特征被暴露，则可能对用户的信息安全产生终身的危害风险。当然，最便捷的模板库保护方法是依赖授权管理者的授权访问控制策略，其安全性是不可靠的。具有很好的安全性保障的保护方法是基于用户的模糊金库方法。模糊金库方法是一种采用数学方式将用户的某种属性集合或生物特征与预先选定的秘密信息绑定在一起的算法，使用户的属性集合或生物特征以及秘密信息均免遭泄露的风险[14]。

上面简述的5级安全技术要求是对一般性的数据库管理系统而言的，但是目前还没有专门针对学术数据库的安全技术要求。2017年6月开始实施的《中华人民共和国网络安全法》第四十五条规定“依法负有网络安全监督管理职责的部门及其工作人员，必须对在履行职责中知悉的个人信息、隐私和商业秘密严格保密，不得泄露、出售或者非法向他人提供”。依据该安全法规定以及科研人员对国家建设与发展的重要性，应该建立相应的针对学术数据库用户的隐私保护的安全技术要求。建议依据科研院校属性与等级、科研人员的职级及其个人隐私信息的重要程度、学术数据库的等级分类，参考《信息安全技术－数据库管理系统安全技术要求》标准，制定相应的以科研用户隐私保护为主导的学术数据库安全技术（管理）标准要求或法规。

3.5 规范数据库信息采集权限

学术数据库的服务商同时也是网络信息服务的提供商，对引进的国外数据库，应该在采购合同中依据《中华人民共和国网络安全法》对用户的隐私保护条款加以细化，规范数据库平台的信息采集权限。条款不仅针对个人用户，对机构及国家因素也应该一并体现。要考虑用户注册数据信息在本地存留与跨境流动上的问题，建议国外数据库在国内设立镜像站，减少或停止使用第三方脚本对用户信息数据的收集，承诺保证用户数据不被共享、滥用或泄露。在2013年“棱镜门”事件中，美国执法机构正是利用了美国企业掌握大量跨境数据源的便利而实施了远远超出处置严重犯罪的维护国家安全范围的个人实时监控和数据调取，并且不受监管约束[15]。因此，国内科研用户的隐私信息保护问题应该引起图书馆管理部门的重视和关注，而且在当前国家领导人多次在重要场合提及网络安全涉及国家安全的情况下，有必要构建相应的以保护科研用户隐私信息安全为目的防止学术数据库外挂采集读者信息的脚本软件等信息自动采集工具的监管机制或系统。

3.6 信息流分散控制、噪声干扰及安全审计策略与隐私保护

在2015年世界著名期刊 *Science* 发表的文章 *The end of privacy* 中，M. Enserink 与 G. Chin 两位作者阐述到[16]，从大数据到无处不在的互联网连接，技术赋予研究人员和公众更强的能力，但却使传统的隐私概念过时了。那么在当前的大数据与互联网时代，公众的隐私权限将如何得到有效的保护？尤其是几乎每天都接触到互联网上重要学术数据库的科研用户。

在大数据环境下，图书馆应适时地采用一些最新的隐私保护技术来防止用户的个人隐私信息遭受侵犯。大数据个人隐私保护主要涉及到数据产生、传输、使用、共享、存储、存档、销毁等方面以及用户的位置信息的隐私保护问题，相关的隐私保护技术也不断出现。如 I. Roy 等[17]提出的基于分散信息流控制和差分隐私保护技术隐私保护系统 Airavat。该系统可以在运行大数据分析计算系统 MapReduce 过程中，阻止未经许可的隐私泄露。针对提供商可能根据用户的需求推断用户的隐私信息的问题，G. Zhang 等提出了一种噪声产生策略 HPNGS[18]，即根据用户需求历史发生的概率产生需求噪声，使得产生的噪声需求与真实需求达到相同的发生概率，这样可使服务提供商很难辨别用户的真实需求，从而达到隐私保护的目的。C. Wang 等提出的利用云存储的隐私保护公共审计策略[19]，此方法使得用户可以借助第三方审计来检测数据的安全性，而审计过程不会对用户数据的隐私产生新的安全风险。

3.7 加密算法与隐私保护

同态加密是一种允许对密文进行某特定代数操作的加密技术，而全同态加密[20-21]则是一种允许对密文进行任何代数操作（通常指加法与乘法运算）的加密技术。运用（全）同态加密技术对加密数据进行代数操作后的结果进行解密后得到的数据结果，与对明文进行相同的代数操作得到的数据结果相同。利用（全）同态加密技术，可以将用户的涉及隐私的信息在需要作大数据分析处理前进行加密运算，这样就可以避免学术数据库及有关的第三方机构在获取用户的个

人阅读或科研兴趣等信息时隔离个人的隐私信息。

3.8 位置匿名算法与隐私保护

目前有许多科研用户常使用手机来浏览检索网络资源或查阅学术数据库，而基于移动定位的技术(GPS、Wifi)等的定位服务 LBSs(Location-based Services)可被数据库机构或第三方服务商用来持续地跟踪用户的位置，根据对用户精确的物理定位为他们提供服务。如开发新的移动应用，提高个性化搜索结果，提供移动广告服务等。利用 LBSs 的应用在给各方带来利益的同时，也泄露或干扰了移动用户的个人生活或隐私信息，如用户可被跟踪而暴露其家庭地址、工作单位、生活规律，或被基于位置的植入广告打扰等。

为了解决用户的位置隐私问题，C. Ardagna 等[22]提出了一种对测量到的位置信息进行模糊处理的解决方案。此后 X. Pan 等[23]为防止基于位置的相关攻击，采用位置 k-匿名和隐身粒度作为隐私指标，提出了一种新的基于团增量的匿名算法 ICliqueCloak。其主要思想是随着手机用户位置的不断更新，及时地产生新的无向图形式的极大匿名位置区域。

3.9 区块链技术与隐私保护

区块链是采用分布式数据存储、点对点传输、共识机制以及密码算法等计算机应用技术有机整合构建的一种新型互联网数据库模式。简单地说，区块链也就是一种分布式账本。区块链的特点是去中心化、公开透明，让每个人均可参与数据库记录。区块链技术是基于密码学原理的一种去中心化或部分去中心化的分布式数据安全记录机制。借助现代密码技术，区块链技术可以带给用户身份的隐私保护及记入数据的机密性保护[24]，并可依据监管要求选择性地进行公开用户的身份或记入的隐私数据。如今，区块链技术已在数字货币、数字票据、支付清算、金融交易、权益证明、证券交易、保险管理、身份证件、金融审计、物联网等领域得到快速发展，应用前景非常广阔。如迪拜计划开发基于区块链技术的电子护照，沃尔玛使用区块链追踪运输无人机，瑞典政府测试利用区块链进行能源交换，麻省理工学院已开始用区块链技术颁发毕业证，还有中国银监会正计划推出区块链证券市场等。在区块链中一般采用 Hash 函数、公钥加密算法等来实现用户身份的保密性及用户隐私信息的安全性。在区块链上，用户的保密性与隐私性安全主要体现在 3 个方面：①未经授权的第三方不能通过记录在区块链上的信息识别出交易方(也可以称为用户)，除非有一个交易方将该信息泄露给第三方；②对任何未经授权的第三方来说，交易细节(也可以是个人的隐私信息)是不可见的，除非有一个交易方选择披露该信息给第三方；③不能通过对交易细节进行核对整理、分析，或者与非区块链上数据进行比对等方法获得有关交易方或交易细节的任何信息。

依据我国最新颁布的自 2017 年 6 月 1 日起施行的《中华人民共和国网络安全法》的规定，网站运营机构必须在提供网络服务过程中，依法留存用户登陆的网络日志。科研用户的个人信息不仅涉及到个人隐私信息，且可能会关系他背后的国家或国防科技进步及战略发展现状与规划，所以针对从事或涉及国家或国防前沿性科学技术研究的科研人员，图书馆行业可借鉴基于区块链技术的个人隐私保护技术来管理他们登陆访问科研资源的网络日志。对引进的国外学术数据库系统及第三方平台进行监控，并定期对其网络日志进行审计。在购买引进时，可协议聘请国内在信息安全领域具有高影响力的机构预先对该数据库是否存在泄露用户隐私的安全漏洞问题进行安全测评，可要求数据库服务商对存在的安全问题进行处理。其后可考虑在国内搭建基于区块链技术的国外数据库的可信检索平台。这样的国外数据库可信检索平台是一个可跨库检索多个异构数据库的统一检索平台，图书馆引进购买的所有国外学术数据库都可以融合到这个平台上。这种检索平台最重要的组成部分是其日志系统是基于区块链技术的，该平台既利用现代密码算法(如最新的环签名算法、安全多方计算、非交互式零知识证明等)来实现用户在网络日志上隐私保护功能，又能满足实名制的审计要求。

参考文献：

[1] 姜山，王刚. 大数据对图书馆的启示[J]. 图书馆工作与研究，2013(4):52－54.

[2] 彭华杰. 大数据时代图书馆读者的隐私危机与隐私保护[J]. 图书馆工作与研究，2014(12):56－59.

[3] TAJANI A[EB/OL]. [2018－03－19]. https://twitter.com/ep_president/status/975683240777453569.

[4] 清华大学数据库导航系统－x server 版[EB/OL]. [2018－11－20]. http://nav.lib.tsinghua.edu.cn/xport/dbdh.htm.

[5] POLANKA S. OverDrive announces a series of "Big Data" reports | NSR[EB/OL]. [2017－03－20]. http://www.noshelfrequired.com/overdrive-announces-a-series-of-big-data-reports/.

[6] IEEE 隐私政策. IEEE-privacy & opting out of cookies [EB/OL]. [2017－04－05]. http://www.ieee.org/about/help/security_privacy.html.

[7] Elsevier privacy policy [EB/OL]. [2016－12－29]. https://www.elsevier.com/legal/privacypolicy.

[8] SpringerLink 隐私声明. Legal information Springer [EB/OL]. [2017-04-05]. https://link. springer. com/termsandconditions.
[9] OverDrive 帐户隐私政策[EB/OL]. [2018-11-20]. http://company. overdrive. com/privacy-policy/.
[10] 廖璠,黄丹珠. 国外图书馆用户隐私保护理论与实践研究进展[J]. 情报理论与实践,2012(4):120-123.
[11] 孟小峰,张啸剑. 大数据隐私管理[J]. 计算机研究与发展,2015(2):265-281.
[12] 李爱国,曹翔,汪社教. 图书馆用户信息资源化过程中用户隐私信息保护问题与对策[J]. 图书情报工作,2015,59(13):26-30.
[13] 中共中央网络安全和信息化领导小组办公室. 国家网络空间安全战略[EB/OL]. [2018-11-20]. http://www. cac. gov. cn/2016-12/27/c_1120195926. htm? from=timeline.
[14] YOU L, YANG L, YU W, et al. A cancelable fuzzy vault algorithm based on transformed fingerprint features[J]. Chinese journal of electronics, 2017, 26(2):236-243.
[15] 张衡,李兆雄. 大数据本地存留与跨境流动问题研究[J]. 信息安全与通信保密,2015(6):65-69.
[16] ENSERINK M, CHIN G. The end of privacy[J]. Science, 2015, 347(6221), 490-491.
[17] ROY I, SETTY S, KILZER A, et al. Airavat: security and privacy for MapReduce[C]//Proceedings of the 7th USENIX conference on networked systems design and implementation. San Jose: USENIX Association Berkeley, 2010:1-16.
[18] ZHANG G, YANG Y, CHEN J. A historical probability based noise generation strategy for privacy protection cloud computation[J]. Journal of computer and system sciences, 2012, 78(5):1374-1381.
[19] WANG C, WANG Q, REN K, et al. Privacy-preserving public auditing for data storage security in cloud computing[C]// Proceedings of the 29th INFOCOM 2010. Piscataway: IEEE, 2010: 1-9.
[20] ZHANG L, YAN Z, KANTOLA R. A review of homomorphic encryption and its applications [C]// Proceedings of the 9th EAI international conference on mobile multimedia communication. Brussels: ICST, 2016: 97-106.
[21] CHILLOTTI I, GAMA N, GEORGIEVA M, et al. Faster fully homomorphic encryption: bootstrapping in less than 0. 1 Seconds [C]// Proceedings of the 22nd international conference on the theory and application of cryptology and information security (ASIACRYPT 2016). Berlin: Springer, 2016: 3-33.
[22] ARDAGNA C, CREMONINI M, SABRINA D, et al. An obfuscation-based approach for protecting location privacy[J]. IEEE transactions on dependable and secure computing, 2010, 8(1): 13-27.
[23] PAN X, XU J, MENG X. Protecting location privacy against location-dependent attacks in mobile services[J]. IEEE transactions on knowledge and data engineering, 2012, 24(8): 1506-1519.
[24] YANG D, GAVIGAN J, O'HEARN Z. Survey of confidentiality and privacy preserving technologies for blockchains [EB/OL]. [2018-04-05]. https://www. r3cev. com/blog/2017/3/8/survey-of-confidentiality-and-privacy-preserving-technologies-for-blockchains.

作者贡献说明:
艾琼:提出研究思路,撰写及修改论文;
刘纯璐:进行数据采集,撰写及修改论文;
游林:确定研究进程,修订论文最后版本。

Privacy Protection and Countermeasures for Scientific Research Users Accessing to Academic Databases in Big Data Environment

Ai Qiong[1] Liu Chunlu[2] You Lin[3]
[1] Library, Hangzhou Dianzi University, Hangzhou 310018
[2] School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018
[3] College of Cyberspace Security, Hangzhou Dianzi University, Hangzhou 310018

Abstract: [**Purpose/significance**] Aiming at the privacy protection issues involved in the process of Chinese scientific research user accessing foreign academic information, this paper puts forward some suggestions, countermeasures and prevention strategies. The paper also studied the protection technologies about the big data and block chain privacy, and feasible privacy protection strategies. [**Method/process**] By collecting and analyzing the network access data of some representative foreign academic database platforms, it is found that not only the platforms themselves insert the scripts on their websites to collect user information, but also there exist real-time collections and analyses of user information using the third-party platforms. [**Result/conclusion**] Under the big data environment, through large data technologies, the information can be used by database vendors and their partners or government agencies, and it will lead to the risk of disclosure of privacy information of scientific research users, which may cause some personal safety issues, personal financial problems, or even national security issues.

Keywords: big data foreign database research users privacy protection block chain